Biostatistics 602 - Statistical Inference
Lecture 15
Bayes Estimator

Hyun Min Kang

March 12th, 2013

Last Lecture

- Can Cramer-Rao bound be used to find the best unbiased estimator for any distribution? If not, in which cases?
- When Cramer-Rao bound is attainable, can Cramer-Rao bound be used for find best unbiased estimator for any $\tau(\theta)$? If not, what is the restriction on $\tau(\theta)$?
- What is another way to find the best unbiased estimator?
- Describe two strategies to obtain the best unbiased estimators for $\tau(\theta)$, using complete sufficient statistics.

Recap - The power of complete sufficient statistics

Theorem 7.3.23

Let T be a complete sufficient statistic for parameter θ. Let $\phi(T)$ be any estimator based on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Finding UMUVE - Method 1

Use Cramer-Rao bound to find the best unbiased estimator for $\tau(\theta)$.

1. If "regularity conditions" are satisfied, then we have a Cramer-Rao bound for unbiased estimators of $\tau(\theta)$.
   - It helps to confirm an estimator is the best unbiased estimator of $\tau(\theta)$ if it happens to attain the CR-bound.
   - If an unbiased estimator of $\tau(\theta)$ has variance greater than the CR-bound, it does NOT mean that it is not the best unbiased estimator.
2. When "regularity conditions" are not satisfied, $\frac{[\tau'(\theta)]^2}{I_n(\theta)}$ is no longer a valid lower bound.
   - There may be unbiased estimators of $\tau(\theta)$ that have variance smaller than $\frac{[\tau'(\theta)]^2}{I_n(\theta)}$.

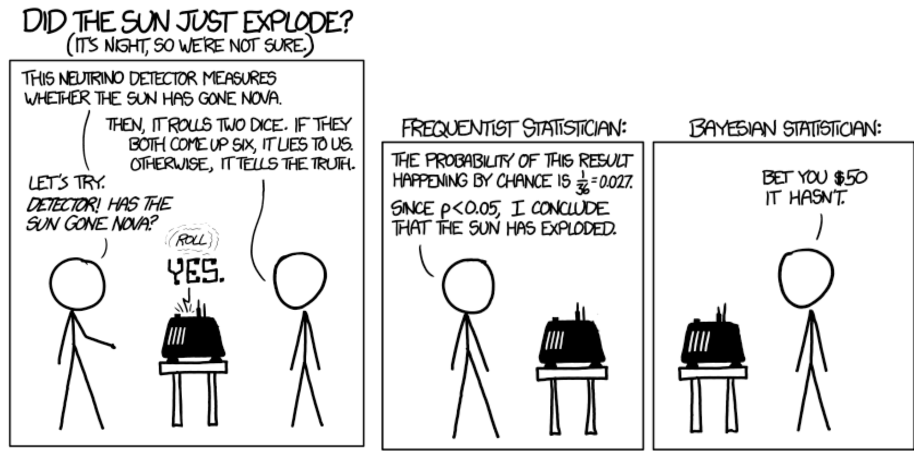
Finding UMVUE - Method 2

Use complete sufficient statistic to find the best unbiased estimator for $\tau(\theta)$.

1. Find complete sufficient statistic T for θ.
2. Obtain $\phi(T)$, an unbiased estimator of $\tau(\theta)$ using either of the following two ways
   - Guess a function $\phi(T)$ such that $\mathrm{E}[\phi(T)] = \tau(\theta)$.
   - Guess an unbiased estimator $h(\mathbf{X})$ of $\tau(\theta)$. Construct $\phi(T) = \mathrm{E}[h(\mathbf{X})|T]$, then $\mathrm{E}[\phi(T)] = \mathrm{E}[h(\mathbf{X})] = \tau(\theta)$.

Frequentists vs. Bayesians

A biased view in favor of Bayesians at http://xkcd.com/1132/



Bayesian Statistic

Frequentist's Framework

$$\mathcal{P} \quad = \quad \{\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta), \theta \in \Omega\}$$

Bayesian Statistic

- Parameter θ is considered as a random quantity
- Distribution of θ can be described by probability distribution, referred to as *prior* distribution
- A sample is taken from a population indexed by θ, and the prior distribution is updated using information from the sample to get *posterior* distribution of θ given the sample.

Bayesian Framework

- Prior distribution of θ : $\theta \sim \pi(\theta)$.
- Sample distribution of $\mathbf{X}$ given θ.
$$\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$$
- Joint distribution $\mathbf{X}$ and θ
$$f(\mathbf{x}, \theta) = \pi(\theta) f(\mathbf{x}|\theta)$$
- Marginal distribution of $\mathbf{X}$.
$$m(\mathbf{x}) \quad = \quad \int_{\theta \in \Omega} f(\mathbf{x}, \theta)\, d\theta = \int_{\theta \in \Omega} f(\mathbf{x}|\theta)\pi(\theta)\, d\theta$$
- Posterior distribution of θ (conditional distribution of θ given $\mathbf{X}$)
$$\pi(\theta|\mathbf{x}) \quad = \quad \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \qquad \text{(Bayes' Rule)}$$

Example

Burglary (θ)	Pr(Alarm|Burglary) $= \Pr(X = 1\|\theta)$
True ($\theta = 1$)	0.95
False ($\theta = 0$)	0.01

Suppose that Burglary is an unobserved parameter ($\theta \in \{0, 1\}$), and Alarm is an observed outcome ($X = \{0, 1\}$).

- Under Frequentist's Framework,
  - If there was no burglary, there is 1% of chance of alarm ringing.
  - If there was a burglary, there is 95% of chance of alarm ringing.
  - One can come up with an estimator on θ, such as MLE
  - However, given that alarm already rang, one cannot calculate the probability of burglary.

Inference Under Bayesian's Framework

Leveraging Prior Information

Suppose that we know that the chance of Burglary per household per night is 10^{-7}.

$$\Pr(\theta = 1|X = 1) = \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(X = 1)} \qquad \text{(Bayes' rule)}$$

$$= \Pr(X = 1|\theta = 1)\frac{\Pr(\theta = 1)}{\Pr(\theta = 1, X = 1) + \Pr(\theta = 0, X = 1)}$$

$$= \frac{\Pr(X = 1|\theta = 1)\Pr(\theta = 1)}{\Pr(X = 1|\theta = 1)\Pr(\theta = 1) + \Pr(X = 1|\theta = 0)\Pr(\theta = 0)}$$

$$= \frac{0.95 \times 10^{-7}}{0.95 \times 10^{-7} + 0.01 \times (1 - 10^{-7})} \approx 9.5 \times 10^{-6}$$

So, even if alarm rang, one can conclude that the burglary is unlikely to happen.

What if the prior information is misleading?

Over-fitting to Prior Information

Suppose that, in fact, a thief found a security breach in my place and planning to break-in either tonight or tomorrow night for sure (with the same probability). Then the correct prior $\Pr(\theta = 1) = 0.5$.

$$\Pr(\theta = 1|X = 1)$$

$$= \frac{\Pr(X = 1|\theta = 1)\Pr(\theta = 1)}{\Pr(X = 1|\theta = 1)\Pr(\theta = 1) + \Pr(X = 1|\theta = 0)\Pr(\theta = 0)}$$

$$= \frac{0.95 \times 0.5}{0.95 \times 0.5 + 0.01 \times (1 - 0.5)} \approx 0.99$$

However, if we relied on the inference based on the incorrect prior, we may end up concluding that there are $> 99.9\%$ chance that this is a false alarm, and ignore it, resulting an exchange of one night of good sleep with quite a bit of fortune.

Advantages and Drawbacks of Bayesian Inference

Advantages over Frequentist's Framework

- Allows making inference on the distribution of θ given data.
- Available information about θ can be utilized.
- Uncertainty and information can be quantified probabilistically.

Drawbacks of Bayesian Inference

- Misleading prior can result in misleading inference.
- Bayesian inference is often (but not always) prone to be "subjective"
  - See : Larry Wasserman "Frequentist Bayes is Objective" (2006) Bayesian Analysis 3:451-456.
- Bayesian inference could be sometimes unnecessarily complicated to interpret, compared to Frequentist's inference.

Bayes Estimator

Definition

Bayes Estimator of θ is defined as the posterior mean of θ.

$$E(\theta|\mathbf{x}) = \int_{\theta\in\Omega} \theta\pi(\theta|\mathbf{x})d\theta$$

Example Problem

$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ where $0 \le p \le 1$. Assume that the prior distribution of p is $\text{Beta}(\alpha, \beta)$. Find the posterior distribution of p and the Bayes estimator of p, assuming α and β are known.

Solution (1/4)

Prior distribution of p is

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Sampling distribution of $\mathbf{X}$ given p is

$$f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{n} \left\{ p^{x_i}(1-p)^{1-x_i} \right\}$$

Joint distribution of $\mathbf{X}$ and p is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}, p) &= f_{\mathbf{X}}(\mathbf{x}|p)\pi(p) \\ &= \prod_{i=1}^{n} \left\{ p^{x_i}(1-p)^{1-x_i} \right\} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \end{aligned}$$

Solution (2/4)

The marginal distribution of $\mathbf{X}$ is

$$\begin{aligned} m(\mathbf{x}) &= \int f(\mathbf{x}, p)\,dp = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum_{i=1}^n x_i+\alpha-1}(1-p)^{n-\sum_{i=1}^n x_i+\beta-1}dp \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\ &\quad \times \frac{\Gamma(\sum x_i+\alpha+n-\sum x_i+\beta)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)} \\ &\quad \times \int_0^1 f_{\text{Beta}(\sum x_i+\alpha, n-\sum x_i+\beta)}(p)\,dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(\alpha+\beta+n)} \end{aligned}$$

Solution (3/4)

The posterior distribution of $\theta|\mathbf{x}$:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}, p)}{m(\mathbf{x})} \\ &= \frac{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1}\right]}{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)}{\Gamma(\alpha+\beta+n)}\right]} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\sum x_i+\alpha)\Gamma(n-\sum x_i+\beta)} p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1} \end{aligned}$$

Solution (4/4)

The Bayes estimator of p is

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^n x_i + \alpha}{\sum_{i=1}^n x_i + \alpha + n - \sum_{i=1}^n x_i + \beta} = \frac{\sum_{i=1}^n x_i + \alpha}{\alpha + \beta + n} \\ &= \frac{\sum_{i=1}^n x_i}{n} \frac{n}{\alpha+\beta+n} + \frac{\alpha}{\alpha+\beta} \frac{\alpha+\beta}{\alpha+\beta+n} \\ &= \text{[Guess about } p \text{ from data]} \cdot \text{weight}_1 \\ &\quad + \text{[Guess about } p \text{ from prior]} \cdot \text{weight}_2\end{aligned}$$

As n increase, $\text{weight}_1 = \frac{n}{\alpha+\beta+n} = \frac{1}{\frac{\alpha+\beta}{n}+1}$ becomes bigger and bigger and approaches to 1. In other words, influence of data is increasing, and the influence of prior knowledge is decreasing.

Is the Bayes estimator unbiased?

$$E\left[\frac{\sum_{i=1}^n + \alpha}{\alpha+\beta+n}\right] = \frac{np+\alpha}{\alpha+\beta+n} \neq p$$

Unless $\frac{\alpha}{\alpha+\beta} = p$.

$$\text{Bias} = \frac{np+\alpha}{\alpha+\beta+n} - p = \frac{\alpha - (\alpha+\beta)p}{\alpha+\beta+n}$$

As n increases, the bias approaches to zero.

Sufficient statistic and posterior distribution

Posterior conditioning on sufficient statistics

If $T(\mathbf{X})$ is a sufficient statistic, then the posterior distribution of θ given $\mathbf{X}$ is the same to the posterior distribution given $T(\mathbf{X})$. In other words,

$$\pi(\theta|\mathbf{x}) = \pi(\theta|T(\mathbf{x}))$$

Conjugate family

Definition 7.2.15

Let $\mathcal{F}$ denote the class of pdfs or pmfs for $f(x|\theta)$. A class Π of prior distributions is a conjugate family of $\mathcal{F}$, if the posterior distribution is the class Π for all $f \in \mathcal{F}$, and all priors in Π, and all $x \in \mathcal{X}$.

Example: Beta-Binomial conjugate

Let

- $X_1, \cdots, X_n | p \sim \text{Binomial}(m, p)$
- $\pi(p) \sim \text{Beta}(\alpha, \beta)$

where m, α, β is known. The posterior distribution is

$$\pi(p|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha, mn - \sum_{i=1}^{n} x_i + \beta\right)$$

Example: Gamma-Poisson conjugate

- $X_1, \cdots, X_n | \lambda \sim \text{Poisson}(\lambda)$
- $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$
- Prior:
$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\lambda/\beta}$$
- Sampling distribution
$$\begin{aligned} \mathbf{X}|\lambda &\overset{\text{i.i.d.}}{\sim} \frac{e^{-\lambda}\lambda^{x}}{x!} \\ f_{\mathbf{X}}(\mathbf{x}|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \end{aligned}$$

Gamma-Poisson conjugate (cont'd)

- Joint distribution of $\mathbf{X}$ and λ.
$$\begin{aligned} f(\mathbf{x}|\lambda)\pi(\lambda) &= \left[\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right] \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \lambda^{\alpha-1} e^{-\lambda/\beta} \\ &= e^{-n\lambda-\lambda/\beta} \lambda^{\sum x_i + \alpha - 1} \frac{1}{\prod_{i=1}^{n} x_i!} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \end{aligned}$$
- Marginal distribution
$$m(\mathbf{x}) = \int f(\mathbf{x}|\lambda)\pi(\lambda) d\lambda$$

Gamma-Poisson conjugate (cont'd)

- Posterior distribution (proportional to the joint distribution)
$$\begin{aligned} \pi(\lambda|\mathbf{x}) &= \frac{f(\mathbf{x}|\lambda)\pi(\lambda)}{m(\mathbf{x})} \\ &= e^{-n\lambda-\lambda/\beta} \lambda^{\sum x_i + \alpha - 1} \frac{1}{\Gamma(\sum x_i + \alpha)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum x_i + \alpha}} \end{aligned}$$

So, the posterior distribution is $\text{Gamma}\left(\sum x_i + \alpha, \left(n + \frac{1}{\beta}\right)^{-1}\right)$.

Example: Normal Bayes Estimators

Let $X \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ^2, τ^2 are all known, the posterior distribution of θ also becomes normal, with mean and variance given by

$$\begin{aligned} \mathrm{E}[\theta|\mathbf{x}] &= \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu \\ \mathrm{Var}(\theta|x) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2} \end{aligned}$$

- The normal family is its own conjugate family.
- The Bayes estimator for θ is a linear combination of the prior and sample means
- As the prior variance τ^2 approaches to infinity, the Bayes estimator tends toward to sample mean
  - As the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information

Summary

Today

- Bayesian Statistics
- Bayes Estimator
- Conjugate family

Next Lecture

- Bayesian Risk Functions
- Consistency